A Literature Summary: Reduce Clever-Hans Effects in Neural Networks

Jim Neuendorf
Explainable AI for Decision Making, SS 2023

Abstract

As artificial intelligence (AI) is becoming more important, so is the need for the explanation of a model's prediction. Some of the current research focuses on neural networks (NN) that produce the correct output for the wrong reasons. That effect is called the Clever Hans (CH). We discuss methods related to explainable AI for reducing these effects and thus making models more robust.

1. Introduction

In some of the many domains AI is used in, for example medical assistance, decision errors are intolerable. Since AI models typically do not achieve 100% accuracy on complex tasks, explainable AI (XAI) is required to establish trust, fairness and transparency by providing answers to *wh*-questions, e.g. "**Wh**y did the model predict something?" Sometimes, explainability is mandatory by law such as in the European Union with its *right to explanation*. Consequently, XAI techniques must applicable to models processing different types of data such as images, text [1]. Furthermore, foundation models are becoming popular so reducing CH effects is a timely concern [2].

2. Overview

In order to detect CH effects XAI can be a helpful tool. Therefore, we first look at "methods to make explanations […] more robust" [3]. We then discuss methods for reducing CH effects using XAI that either (1) retrain the model using modified data and/or adjusted objectives or (2) change the model structurally, e.g. by inserting layers. The latter is called *post-hoc* because the model is already trained [1].

All papers except the one on anomaly detection focus on supervised learning.

3. Robust Explanations

Explanations can be made more robust using three simple techniques. These are derived from the finding that the maximum possible explanation change is bounded by the Frobenius norm of the Hessian $\|H\|_F$ [3] which the following minimize.

3.1. Curvature Minimization

Here, we penalize $\|H\|_F$ by adding a term to the loss function. Because "calculating the Frobenius norm of the Hessian es expensive, […] [they] propose to estimate the Frobenius norm stochastically" [3] with an expectation value using Monte-Carlo sampling.

3.2. Weight Decay

Because $\|H\|_F$ also depends on the weights of the NN, weight decay can be used to robustify the explanation. This method is a regularization aiming for small weight values. In this case using the Frobenius norm – other regularizations such as L^2 could work as well. Weight decay is a technique for improving model generalization, but in the context of XAI it is new.

3.3. Smooth Activation Functions

Another approach is using activation functions with smaller maximum values of the first and second derivatives. This means the functions is *smooth* which leads to smaller values for $\|H\|_F$. Softplus is such a function. Effects on the model performance are not presented (e.g. in constrast to ReLU).

3.4. Results

The three presented techniques are validated independently using a convolutional neural network (CNN) on the CIFAR-10 dataset. Even small curvature minimizations lead to a significant improvement of the explanation robustness. However, the model accuracy decreases with better the robustness – there is a trade-off.

4. Reducing Clever-Hans Effects

4.1. Bagging in Anomaly Detection

Research on "whether Clever Hans also occurs in unsupervised learning, and in which form, has received so far alsmost no attention." [4] So far only model-specific approaches have been proposed, i.e. no general technique that applies to all anomaly detection models. Anomaly detection models can be categorized into *density-*, *reconstruction-*, and *boundary-based*.

The researchers "introduce a common XAI framework that is applicable to a broad range of anomaly detection models" [4]: They use a three-layer NN architecture for explanation extraction using Deep Tayler Decomposition (a mechanism similar to Layer-wise Relevance Propagation (LRP)). For each of the three layers *feature extraction*, *distance*, and *pooling* they define (1) specific calculations and (2) certain propagation rules – both depending on the model type[1].

For example, the distance (forward pass) of a reconstruction-based model is the Mean Squared Error (MSE), its "outlier score", $o(x) = \|r(x) - x\|^2$. For the attribution (backward pass) in the distance layer the mean μ_{jk} is actually used as opposed to being constant for the other model types.

4.1.1. Evaluation

Using two datasets, both coming with anomaly-ground-truth data, and a Clever-Hans score[2], they find that the different "models are affected by the problem in different ways" [4]. Thus,

[1]The propagation rules (backward pass) are model-dependent because the calculation (forward pass) also is.
[2]The score measures the mismatch between the detection and explanation accuracy.

CH effects "are inherent to the structre of the anomaly detection models rather than […] the trainig data" [4]. This hypothesis is followed by a reasoning about why this is the case.

Sticking with the autoencoder example, they argue that the CH effect is caused by samples whose reconstruction is far away from the input distribution. Thus, an outlier is detected due to features that do not relate to the input distribution which means the outlier detection is unrelated to the actual anomaly.

4.1.2. Solution

Because the source of CH is not in the data but in in the models themselves, the idea is to allow "multiple anomaly models to mutually cancel their individual structural weaknesses" [4]. They propose a bagging approach of outlier scores[3]:

$$o_{\text{Bag}}(x) = \frac{1}{3}\Big(o_{\text{KDE}}(x) + o_{\text{Auto}}(x) + o_{\text{Deep}}(x)\Big)$$

"The bagged model ranks first among all four models […] although relatively far from the ground-truth" [4]. There is still room for improvement by going "beyond simple bagging […] and structurally less rigid models" [4].

4.2. Pruning in Deep Models

CH effects can remain undetected even if the user's explanation agrees with the one from XAI. In this case, there is "neither prior knowledge about the spurious feature, nor data containing it" [2]: We receive the trained model which "is to be robustified post-hoc with limited data" [2] without CH features, apply the soft-pruning, and should be able to deploy the model.

The proposed pruning method is called *Explanation-Guided Exposure Minimization* (EGEM). They arrive at a practical formulation which comprises (1) explanations of the refined model to be close to those of the original one and (2) a penalty used for exposure minimization. The pruning strength depends on each neuron's activation frequency and magnitude.

A second pruning approach *PCA-EGEM* is introduced for more effective pruning: In PCA space the features are more disentangled so pruning should be more effective.

4.2.1. Evaluation

In their work the scientists could validate approve the effectiveness of the pruning method using datasets were the spurious artifacts were known: MNIST with manual poisoning, and ImageNet and ISIC were previous work has identified (potential) CH correlations.

Afterwards, the approach was tested in an exploratory manner on the CelebA dataset. Using PCA-EGEM it could be observed that "a model bias […] has been mitigated" and it "enables the retrieval of a more diverse set of positive instances from a large heterogeneous dataset" [2].

4.3. Class Artifact Compensation in Deep Models

Another unlearning approach using XAI called *Class Artifact Compensation* (ClArC) is provided by [5]. This work focuses on automating the process of CH removal because with datasets getting larger, manual inspection becomes infeasible.

[3]In the equation KDE is a density-, Auto a boundary-, and Deep a boundary-based model.

4.3.1. Large-scale Analysis

For semi-automated discovery they use *Spectral Relevance Analysis* (SpRAy) which aims to bridge the gap between local and global XAI by inspecting large sets of local explanations. From attribution maps it computes a spectrum of eigenvalues via spectral clusters. That spectrum can be used "for ranking […] analyzed classes w.r.t. their potential for exhibiting CH phenomena" [5].

They enhance SpRAy's cluster visualization and labeling by using an intermediate result and propose using Fisher Discriminant Analysis[4] (FDA) to rank class-wise clusterings by CH likeliness, especially with many-class datasets [5].

4.3.2. Unlearning with Augmentative Class Artifact Compensation

Here (A-ClArC), the training data is augmented so that the trained classifier becomes insensitive to an artifact. This is achieved by adding an artifact found with SpRAy to some samples of all other classes using a *forward artifact model*.

When the classifier is trained with the augmented data, it can no longer rely on the artifact as an "easy shortcut" for classification and must learn "real" strategies.

4.3.3. Unlearning with Projective Class Artifact Compensation

The second approach (P-ClArC) suppresses CH artifacts "without retraining by incorporating a *backward artifact model* […] directly into the prediction model" [5]. While the idea is the same as for A-ClArC, data points are projected to a position of the decision boundary to which the estimated artifact direction is normal. The boundary ignores the artifact while leaving clean samples' output unchanged.

4.3.4. Evaluation

The paper contains a very thorough evaluation of the extended SpRAy and ClArC methods on six datasets. They find that their extended SpRAy (1) picks up most artifacts when there are any, even though their importance varies between models and classes, but (2) still requires human judgement for the final decision on CH candidates.

Concerning their proposed methods they conclude that both A-ClArC and P-ClArC perform very well – both in input and latent space. Two limitations are that A-ClArC is time-consuming and "P-ClArC will not lead to an increased generalization performance, since the model never has a chance to adapt its weights […] and correct its faulty prediction" [5].

This way, common large datasets like ImageNet can be "un-Hansed" to provide a more unbiased basis for foundation models.

4.4. Post-hoc Explanations for Unknown Spurious Correlation

The post-hoc approaches discussed so far all relied on XAI and seemed very promising. In the last paper of this summary, the "post-hoc explanation methods tested are ineffective when the

[4]FDA maximizes between-class and minimizes within-class scatter. Other separation algorithms are possible.

spurious artifact is unknwon at test-time" [6]. The used explanations methods are (1) feature attribution, (2) concept activation, and (3) training point ranking.

In order to compare these methods, they define a *spurious score* which quantifies "the strength of a model's dependence on a training signal" [3].

They use three measures comparing different combinations of explanations[5] with three similarity functions (one for each attribution method). Their findings range from (1) struggling to detect signals even when known over (2) "can help detect reliance on the visible signals but not non-visible" to (3) "all methods struggle to reliably indicate that spurious models are reliant on the blur signal" [3].

5. Summary

With the last paper dulling the results outlined previously, it will be interesting to see how the big picture evolves, e.g. how P-ClArC performs across different models and tasks when evaluated in the spurious-score settings. Additionally, further research on how the structure of a model affects explanations and their robustness in supervised learning could impact the development of future foundation models[6].

References

[1] P. Gohel, P. Singh, and M. Mohanty, "Explainable AI: Current status and future directions," *CoRR*, 2021, doi: 10.48550/ARXIV.2107.07045.

[2] L. Linhardt, K.-R. Müller, and G. Montavon, "Preemptively pruning clever-hans strategies in deep neural networks," 2023.

[3] A.-K. Dombrowski, C. J. Anders, K.-R. Müller, and P. Kessel, "Towards robust explanations for deep neural networks," *Pattern Recognit.*, vol. 121, p. 108194, 2022.

[4] J. Kauffmann, L. Ruff, G. Montavon, and K.-R. Müller, "The clever hans effect in anomaly detection," *CoRR*, 2020, doi: 10.48550/ARXIV.2006.10609.

[5] C. J. Anders, L. Weber, et al., "Finding and removing clever hans: Using explanation methods to debug and improve deep models," *Inf. Fusion*, vol. 77, pp. 261–295, 2020.

[6] J. Adebayo, M. Muelly, H. Abelson, and B. Kim, Post hoc explanations may be ineffective for detecting unknown spurious correlation. Presented at Internation Conf. Learn. Representations, 2022.

[5] (1) Known spurious signal detection $S\big(E_{f_{spu}}(x_{spu}), x_{gt}\big)$, (2) cause-for-concern $S\big(E_{f_{spu}}(x_{norm}), E_{f_{norm}}(x_{norm})\big)$, and (3) false alarm $S\big(E_{f_{norm}}(x_{spu}), E_{f_{spu}}(x_{spu})\big)$, where S: similarity, $E_{f_{spu}}$: spurious-model expl., $E_{f_{norm}}$: normal-model expl., x_{gt}: ground-truth expl., x_{norm}: normal input, and x_{spu}: spurious input.

[6] DINOv2 has amazing emergent properties concerning the latent-space structure.